Publishing Your Tools in Your Own Public Galaxy Server

Biology Centre CAS

Laboratory of Molecular cytogenetics

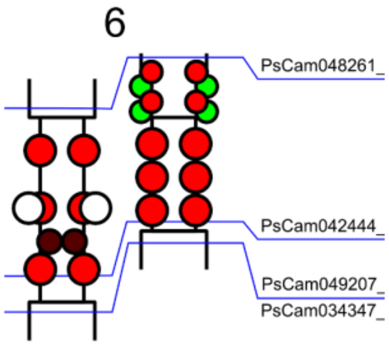
Petr Novak

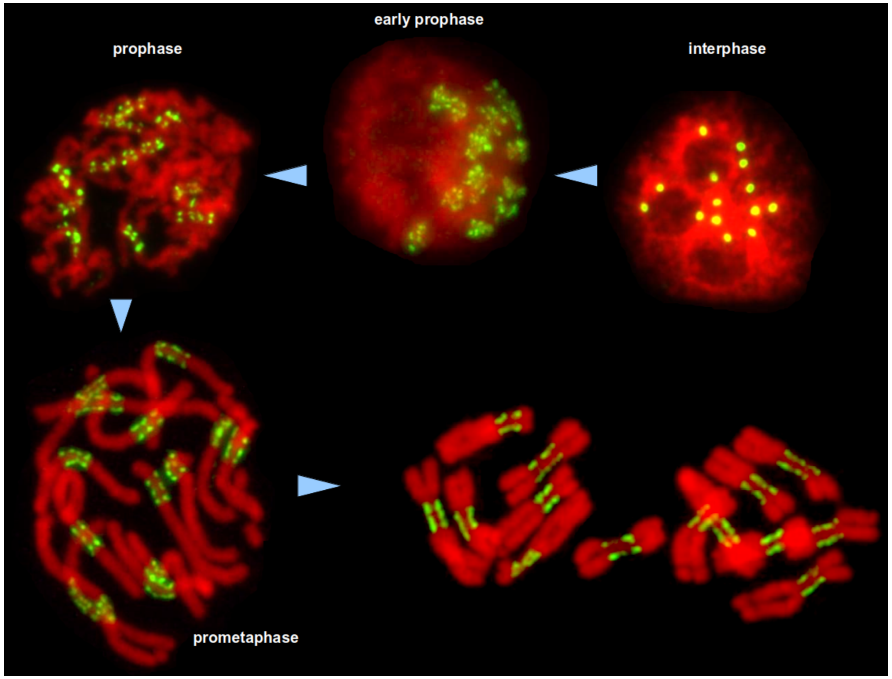
elixir
CZECH
REPUBLIC

Plant Cytogenetics and Genomics

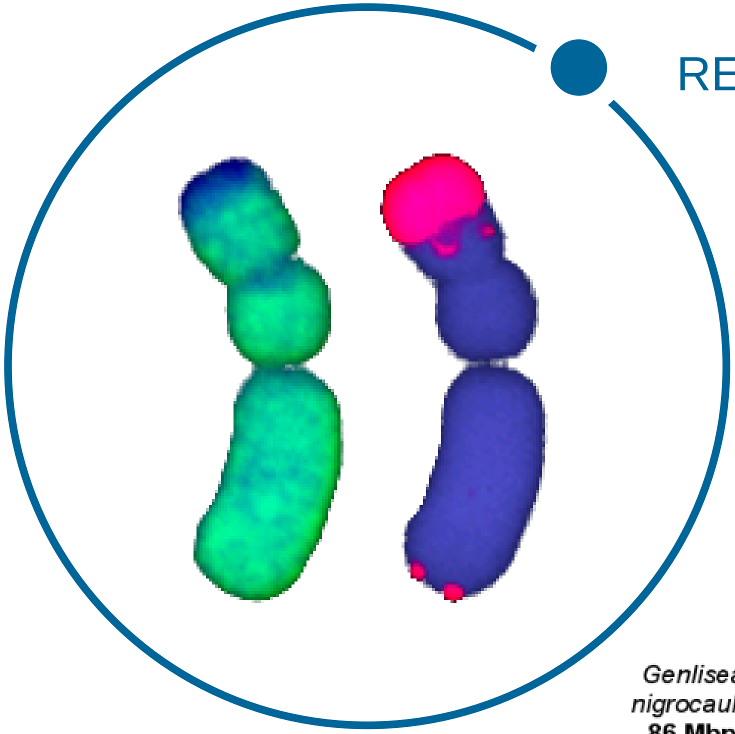
We investigate

- sequence composition and evolution of plant genomes
- chromosome biology and epigenetics
- centromeres and kinetochore





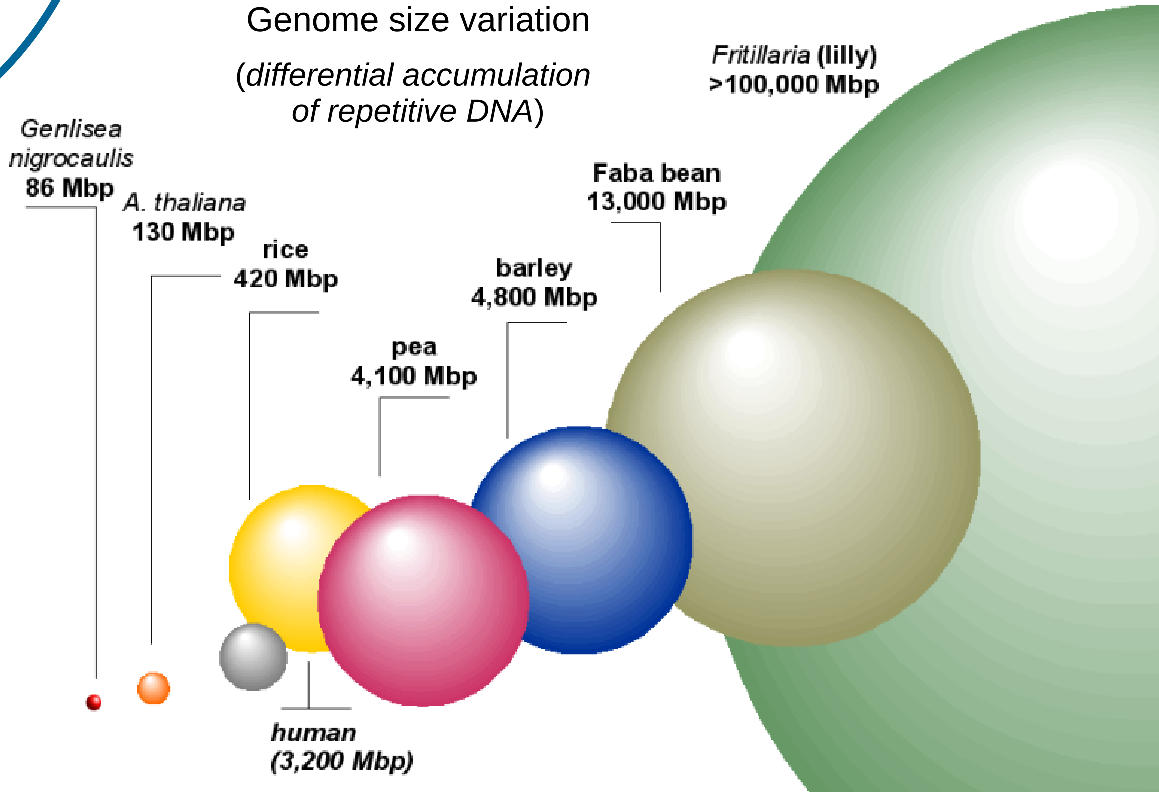
Motivation



REPEATS IN PLANT GENOMES

🟥 Satellite DNA

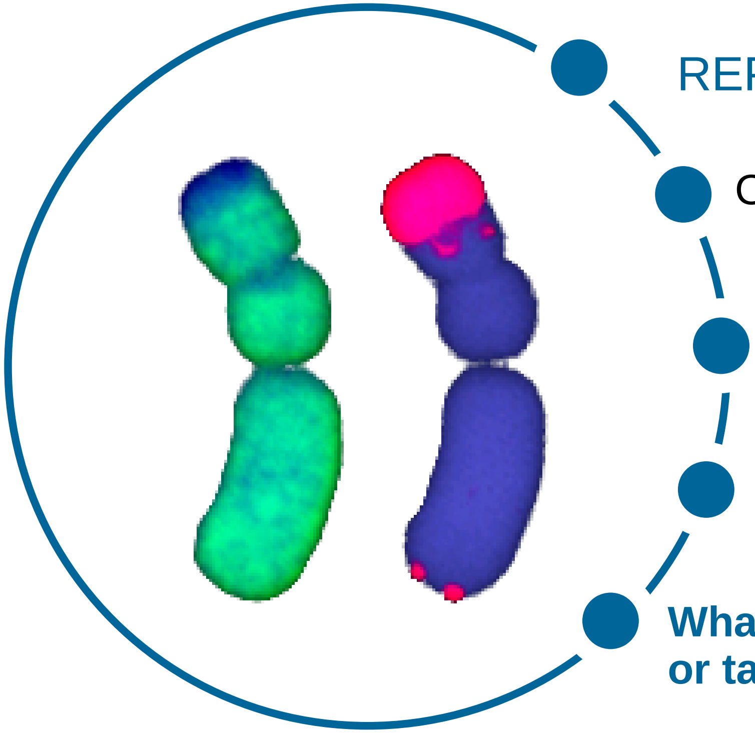
🟩 Retrotransposon



Genome size variation

(*differential accumulation of repetitive DNA*)

Motivation



REPEATS IN PLANT GENOMES

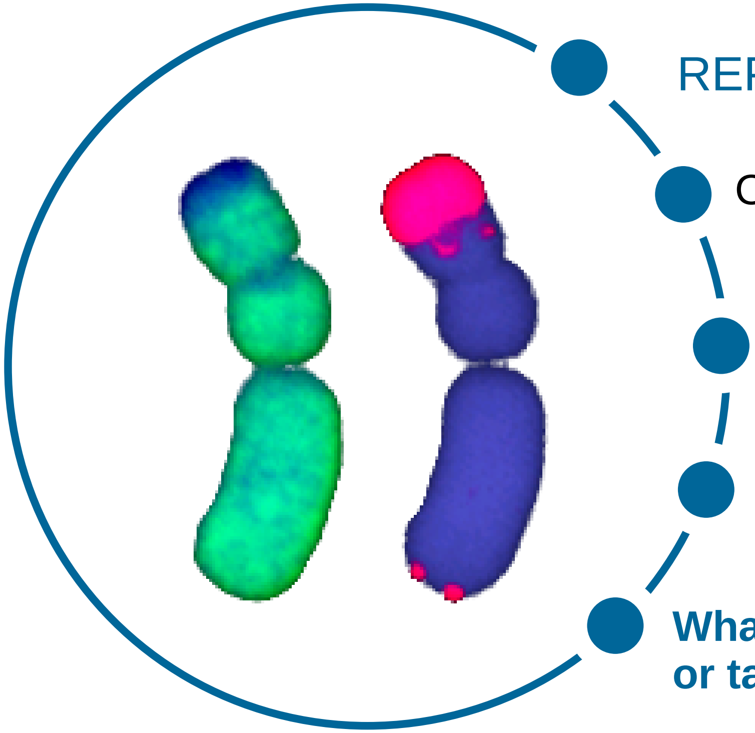
Origin ?

Abundance ?

Evolution ?

What is the repeat composition of individual species or taxa ?

Motivation



REPEATS IN PLANT GENOMES

Origin ?

Abundance ?

Evolution ?

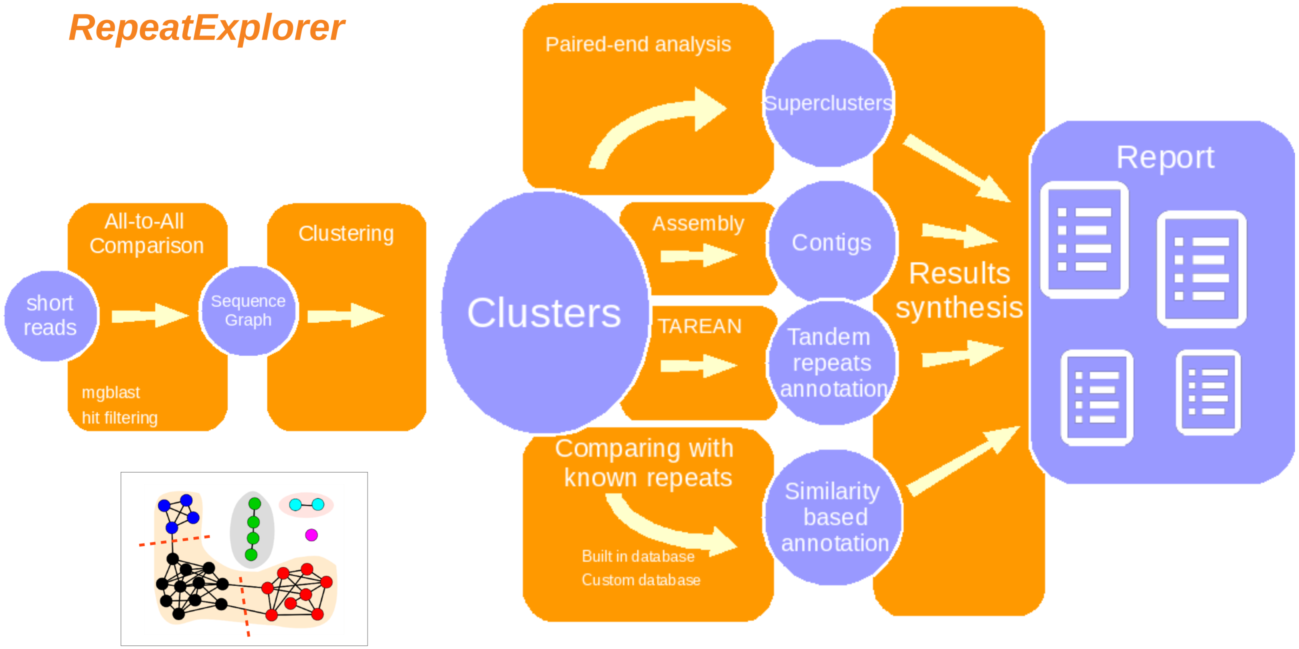
What is the repeat composition of individual species or taxa ?

Introduction of next generation sequencing in ~2005:
→ getting sequence data was no longer a limiting factor …

… but there were no computational tools for repetitive DNA analysis from short reads

RepeatExplorer pipeline

- Unique principle of repeat identification from low-pass WGS data (graph-based clustering)
- *De novo* identification of repeats, no reference DBs, any genome
- Works with short sequence reads (100 nt), no assembly
- Repeat annotation in subsequent steps
- Additional tools



RepeatExplorer

short reads

All-to-All Comparison

mgblast

hit filtering

Sequence Graph

Clustering

Clusters

Paired-end analysis

Superclusters

Assembly

Contigs

TAREAN

Tandem repeats annotation

Comparing with known repeats

Built in database

Custom database

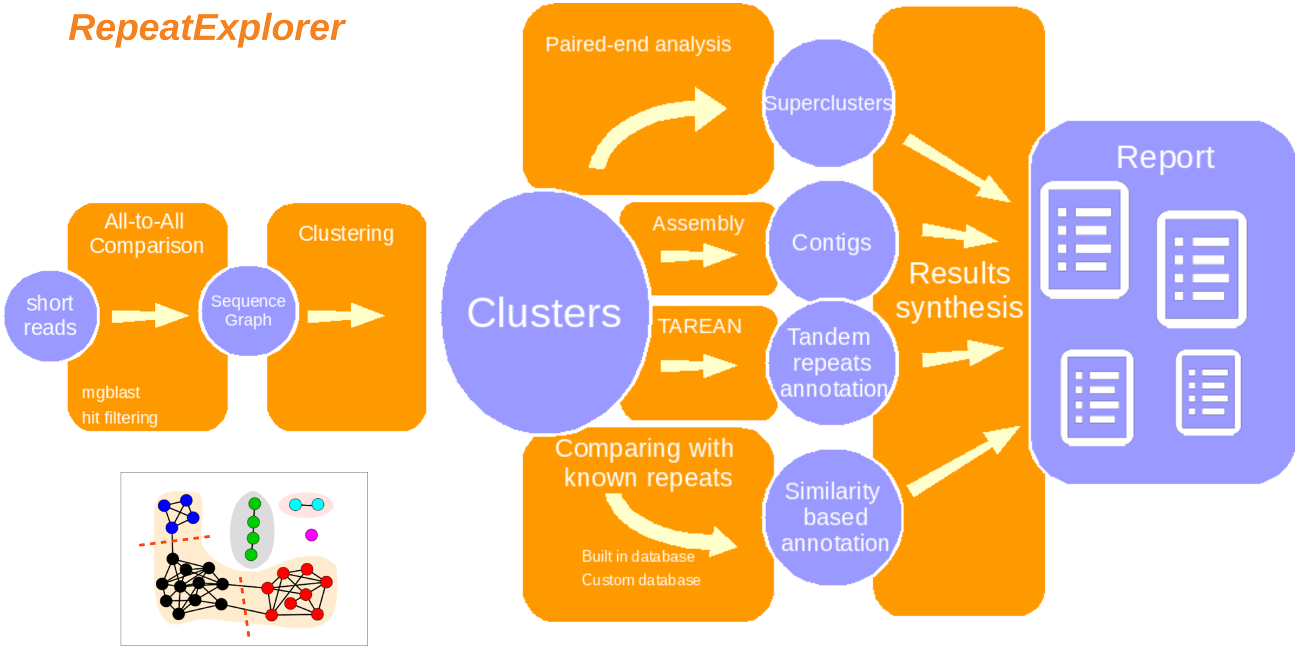
Similarity based annotation

Results synthesis

Report

History

2007	…	2010	…	2013	2014	…	2016	…	2018	2019
• First paper on repeat clustering from NGS data (Macas et al. 2007)		• Introduction of **graph-based clustering** (Novak et al. 2010) ***command-line version***		• **RepeatExplorer in Galaxy** (Novak et al. 2013)						



History

2007 … 2010 … 2013 2014 … 2016 … 2018 2019

- First paper on repeat clustering from NGS data (Macas et al. 2007)

- Introduction of **graph-based clustering** (Novak et al. 2010)

command-line version

- **RepeatExplorer in Galaxy** (Novak et al. 2013)

Why we choose Galaxy platform?

- ***We need to provide are tools to user without access to suitable hardware***

- ***Original pipeline was difficult to setup (our user are usually biologists without experience with computers, Linux,…)***

- ***We needed a job management/scheduling to efficiently utilize our HW***

- ***We needed some kind of user management (Data and job quotas)***

- ***We wanted to reuse our old PBS cluster***

- ***Good documentation***

History

2007 … 2010 … 2013 2014 … 2016 … 2018 2019

- First paper on repeat clustering from NGS data (Macas et al. 2007)

- Introduction of **graph-based clustering** (Novak et al. 2010)

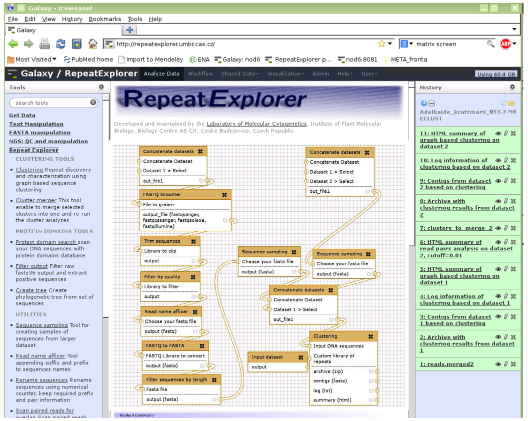
command-line version

- **RepeatExplorer in Galaxy** (Novak et al. 2013)

Public web-based server
(cluster of 12 nodes)

First setup:

- ***Tool definition***
- ***Tool requirement definition***
- ***PBS cluster setup***
- ***NFS storage setup***
- ***Old 'desktops'***



Developed and maintained by the Laboratory of Molecular Cytogenetics, Institute of Plant Molecular Biology, Biology Centre AS CR, Ceske Budejovice, Czech Republic

CLUSTERING TOOLS

- Clustering Repeat discovery and characterization using graph based sequence clustering
- Cluster merger This tool enable to merge selected clusters into one and re-run the cluster analyzes

PROTEIN DOMAINS TOOLS

- Protein domain search scan your DNA sequences with protein domains database
- Filter output Filter raw fasty36 output and extract positive sequences
- Create tree Create phylogenetic tree from set of sequences

UTILITIES

- Sequence sampling Tool for creating samples of sequences from larger dataset
- Read name affixer Tool appending suffix and prefix to sequences names
- Rename sequences Rename sequences using numerical counter, keep required prefix and pair information
- Scan paired reads for overlap Scan paired reads

Adelhaide_kratzmarii_R ECLUST 53.3 MB

11: HTML summary of graph based clustering on dataset 2
10: Log information of clustering based on dataset 2
9: Contigs from dataset 2 based on clustering
8: Archive with clustering results from dataset 2
7: clusters_to_merge_2
6: HTML summary of read pairs analysis on dataset 2, cutoff=0.01
5: HTML summary of graph based clustering on dataset 1
4: Log information of clustering based on dataset 1
3: Contigs from dataset 1 based on clustering
2: Archive with clustering results from dataset 1
1: reads.merged2

History

2007 … 2010 … 2013 2014 … 2016 … 2018 2019

- First paper on repeat clustering from NGS data (Macas et al. 2007)

- Introduction of **graph-based clustering** (Novak et al. 2010)

command-line version

- **RepeatExplorer in Galaxy** (Novak et al. 2013)

Public web-based server
(cluster of 12 nodes)

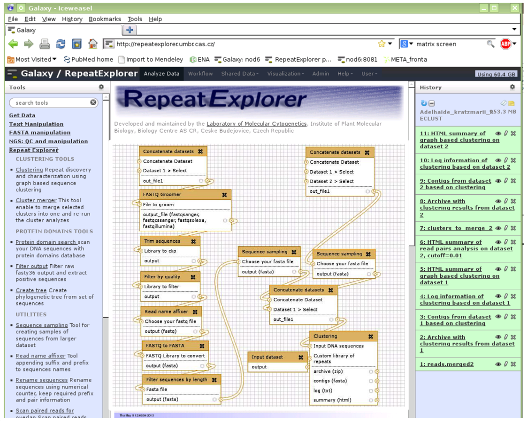
elixir CZECH REPUBLIC

Additional tools

- **TAREAN**
- **REXdb database**
- **ChIP-seq Mapper**
- **ProfRep**
- **DANTE**

Elixir PROJECT $$$

Server migration to CERIT (start)



Developed and maintained by the Laboratory of Molecular Cytogenetics, Institute of Plant Molecular Biology, Biology Centre AS CR, Ceske Budejovice, Czech Republic

- Clustering Repeat discovery and characterization using graph based sequence clustering
- Cluster merger This tool enable to merge selected clusters into one and re-run the cluster analyzes
- Protein domain search scan your DNA sequences with protein domains database
- Filter output Filter raw fasty36 output and extract positive sequences
- Create tree Create phylogenetic tree from set of sequences
- Sequence sampling Tool for creating samples of sequences from larger dataset
- Read name affixer Tool appending suffix and prefix to sequences names
- Rename sequences Rename sequences using numerical counter, keep required prefix and pair information
- Scan paired reads for overlap Scan paired reads

11: HTML summary of graph based clustering on dataset 2
10: Log information of clustering based on dataset 2
9: Contigs from dataset 2 based on clustering
8: Archive with clustering results from dataset 2
7: clusters_to_merge_2
6: HTML summary of read pairs analysis on dataset 2, cutoff=0.01
5: HTML summary of graph based clustering on dataset 1
4: Log information of clustering based on dataset 1
3: Contigs from dataset 1 based on clustering
2: Archive with clustering results from dataset 1
1: reads.merged2

History

2007 … **2010** … **2013** **2014** … **2016** … **2018** **2019**

- First paper on repeat clustering from NGS data (Macas et al. 2007)

- Introduction of **graph-based clustering** (Novak et al. 2010)

command-line version

- **RepeatExplorer in Galaxy** (Novak et al. 2013)

Public web-based server
(cluster of 12 nodes)

elixir CZECH REPUBLIC

Additional tools

- **TAREAN**
- **REXdb database**

- **ChIP-seq Mapper**
- **ProfRep**
- **DANTE**

Elixir PROJECT $$$

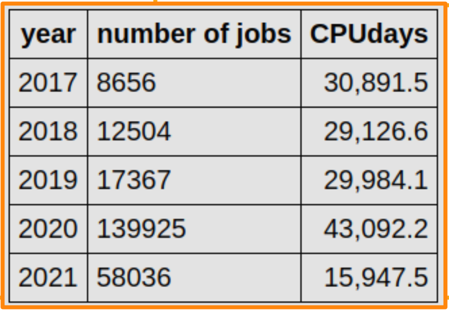
Server migration to CERIT (start)

Server in full use (+ data storage)

Better Hardware

Flexible

Administered by IT proffesionals



year	number of jobs	CPUdays
2017	8656	30,891.5
2018	12504	29,126.6
2019	17367	29,984.1
2020	139925	43,092.2
2021	58036	15,947.5

Frequently used and cited:

RepeatExplorer principle (BMC bioinformatics, 2010)	249 x
RepeatExplore Galaxy server paper (Bioinformatics 2013)	307 x
TAREAN - New tool on Galaxy server (NAR, 2017)	75 x
REXdb database – (Mobile DNA, 2019)	50 x
Nature Protocols (2020)– Galaxy oriented protocols ho to use RE server	

History

2007 … 2010 … 2013 2014 … 2016 … 2018 2019

2007

- First paper on repeat clustering from NGS data (Macas et al. 2007)

2010

- Introduction of **graph-based clustering** (Novak et al. 2010)

command-line version

2013

- **RepeatExplorer in Galaxy** (Novak et al. 2013)

Public web-based server
(cluster of 12 nodes)

2014

elixir CZECH REPUBLIC

2016–2019

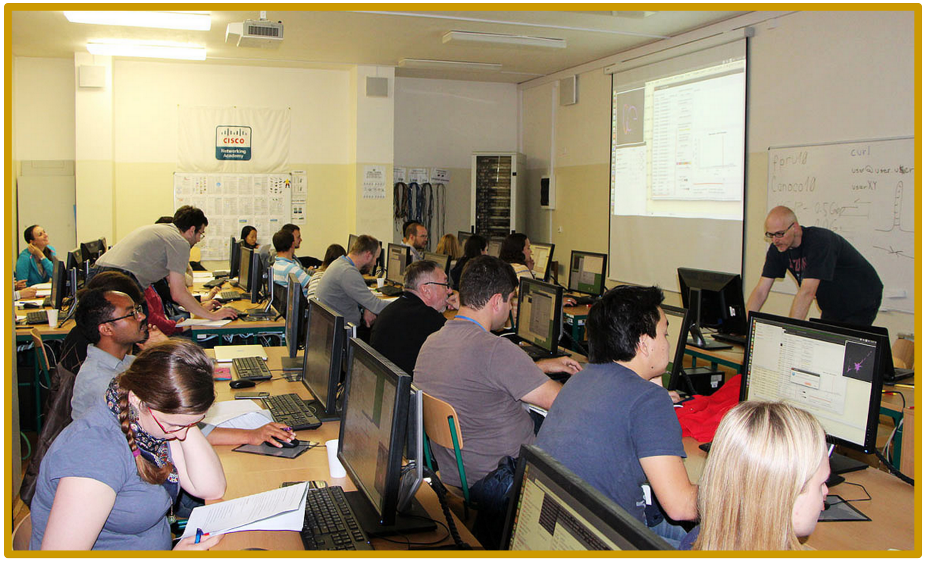
Additional tools

- **TAREAN**
- **REXdb database**
- **ChIP-seq Mapper**
- **ProfRep**
- **DANTE**

Elixir PROJECT $$$

Server migration to CERIT (start)

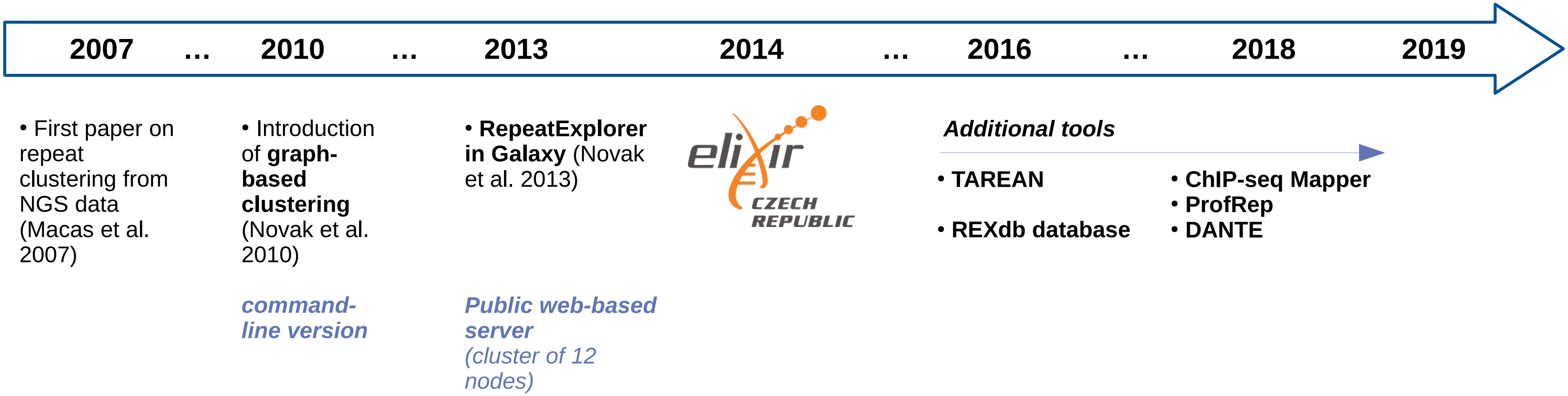
Server in full use (+ data storage)



Training – annual practical workshops – Galaxy oriented

- ~ 40 participants
- 60-70% foreigners, often PhD students
- 3 days, mini-conference + practical training

History



2007 … 2010 … 2013 2014 … 2016 … 2018 2019

2007
- First paper on repeat clustering from NGS data (Macas et al. 2007)

2010
- Introduction of **graph-based clustering** (Novak et al. 2010)

command-line version

2013
- **RepeatExplorer in Galaxy** (Novak et al. 2013)

Public web-based server
(cluster of 12 nodes)

2014

elixir CZECH REPUBLIC

2016 … 2018

Additional tools

- **TAREAN**
- **REXdb database**
- **ChIP-seq Mapper**
- **ProfRep**
- **DANTE**

Additional Benefits of Having RepeatExplorer on Galaxy Server:

- Tool integration, interoperability
- Utilization of RepeatExplorer unrelated tools
  - Data pre-processing
  - Visualization
  - Genome browser integration
- Workflows, protocols sharing, data sharing
- Reproducibility
- Bug reports
- Lower barrier for less experienced users

Repetitive DNA characterization using RepeatExplorer

Plants

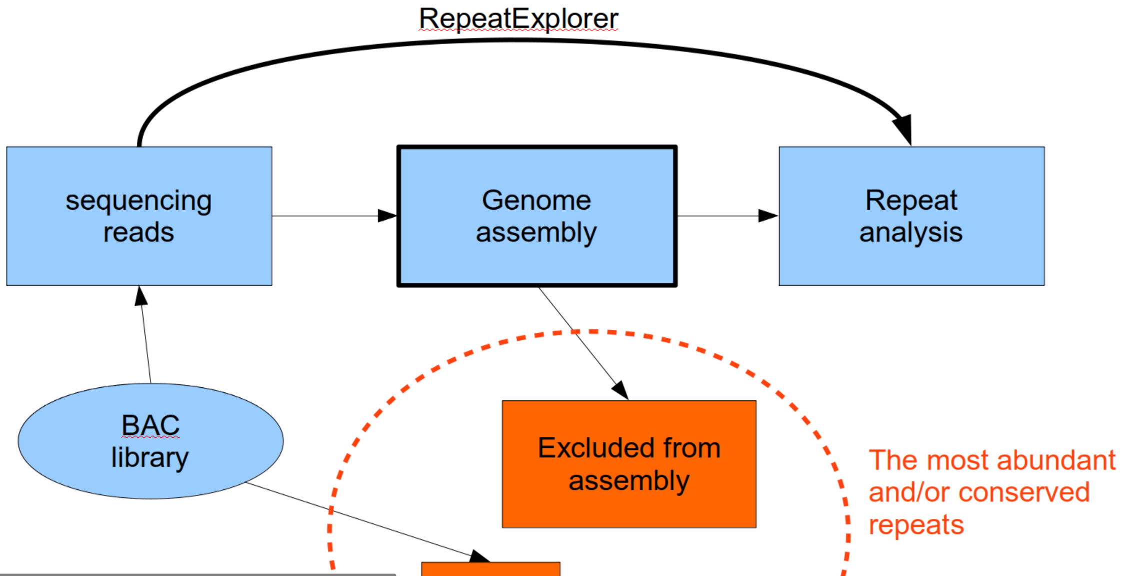
- Over 100 species characterized so far
- Comparative studies
- Whole genome assembly projects

mostly non-model species

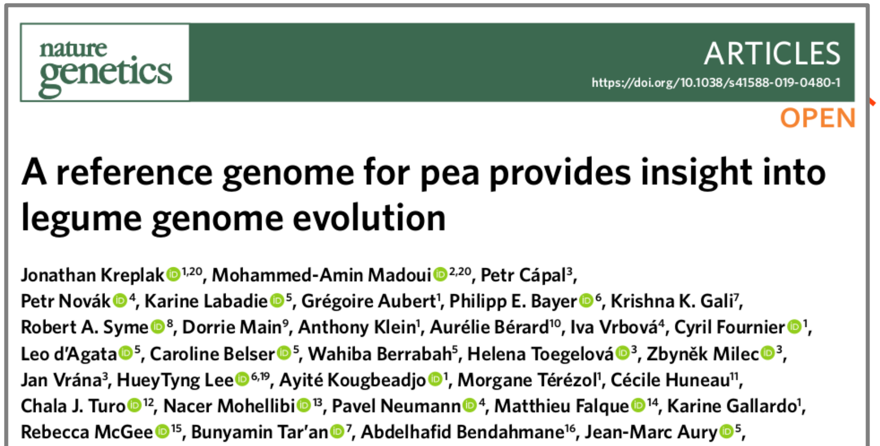
Repetitive DNA characterization using RepeatExplorer

Plants

- Over 100 species characterized so far
- Comparative studies
- Whole genome assembly projects



The most abundant and/or conserved repeats



nature genetics

ARTICLES

https://doi.org/10.1038/s41588-019-0480-1

OPEN

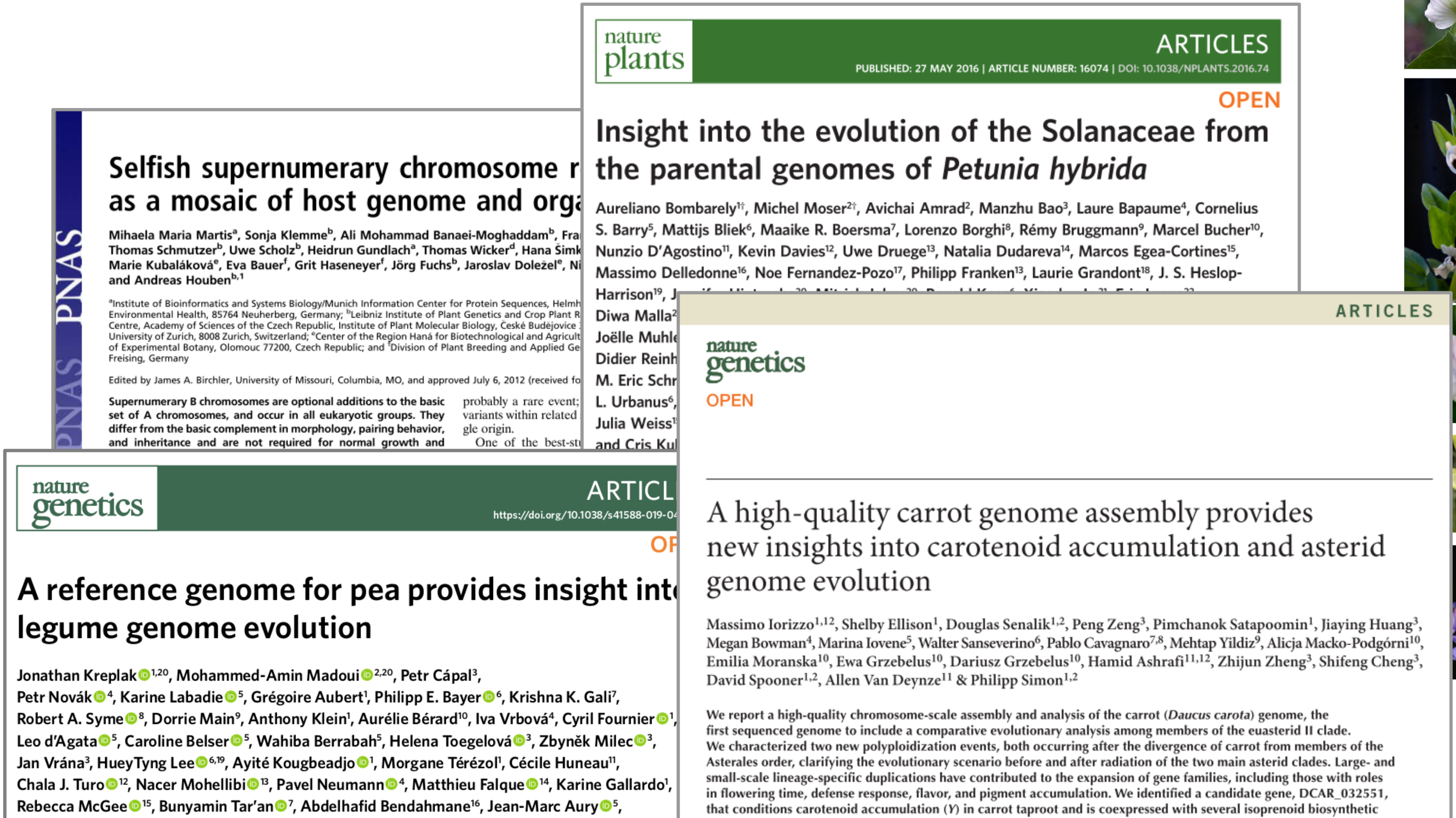
A reference genome for pea provides insight into legume genome evolution

Jonathan Kreplak[1,20], Mohammed-Amin Madoui[2,20], Petr Cápal[3], Petr Novák[4], Karine Labadie[5], Grégoire Aubert[1], Philipp E. Bayer[6], Krishna K. Gali[7], Robert A. Syme[8], Dorrie Main[9], Anthony Klein[1], Aurélie Bérard[10], Iva Vrbová[4], Cyril Fournier[1], Leo d'Agata[5], Caroline Belser[5], Wahiba Berrabah[5], Helena Toegelová[3], Zbyněk Milec[3], Jan Vrána[3], HueyTyng Lee[6,19], Ayité Kougbeadjo[1], Morgane Térézol[1], Cécile Huneau[11], Chala J. Turo[12], Nacer Mohellibi[13], Pavel Neumann[4], Matthieu Falque[14], Karine Gallardo[1], Rebecca McGee[15], Bunyamin Tar'an[7], Abdelhafid Bendahmane[16], Jean-Marc Aury[5],

Repetitive DNA characterization using RepeatExplorer

Plants

- Over 100 species characterized so far
- Comparative studies
- Whole genome assembly projects



nature plants

ARTICLES

PUBLISHED: 27 MAY 2016 | ARTICLE NUMBER: 16074 | DOI: 10.1038/NPLANTS.2016.74

OPEN

Insight into the evolution of the Solanaceae from the parental genomes of *Petunia hybrida*

Aureliano Bombarely[1†], Michel Moser[2†], Avichai Amrad[2], Manzhu Bao[3], Laure Bapaume[4], Cornelius S. Barry[5], Mattijs Bliek[6], Maaike R. Boersma[7], Lorenzo Borghi[8], Rémy Bruggmann[9], Marcel Bucher[10], Nunzio D'Agostino[11], Kevin Davies[12], Uwe Druege[13], Natalia Dudareva[14], Marcos Egea-Cortines[15], Massimo Delledonne[16], Noe Fernandez-Pozo[17], Philipp Franken[13], Laurie Grandont[18], J. S. Heslop-Harrison[19], J

Diwa Malla

Joëlle Muhle

Didier Reinh

M. Eric Schr

L. Urbanus[6],

Julia Weiss

and Cris Kul

PNAS PNAS PNAS

Selfish supernumerary chromosome r

as a mosaic of host genome and orga

Mihaela Maria Martis[a], Sonja Klemme[b], Ali Mohammad Banaei-Moghaddam[b], Fra

Thomas Schmutzer[b], Uwe Scholz[b], Heidrun Gundlach[a], Thomas Wicker[d], Hana Šimk

Marie Kubaláková[e], Eva Bauer[f], Grit Haseneyer[f], Jörg Fuchs[b], Jaroslav Doležel[e], Ni

and Andreas Houben[b,1]

[a]Institute of Bioinformatics and Systems Biology/Munich Information Center for Protein Sequences, Helmh Environmental Health, 85764 Neuherberg, Germany; [b]Leibniz Institute of Plant Genetics and Crop Plant R Centre, Academy of Sciences of the Czech Republic, Institute of Plant Molecular Biology, České Budějovice University of Zurich, 8008 Zurich, Switzerland; [e]Center of the Region Haná for Biotechnological and Agricult of Experimental Botany, Olomouc 77200, Czech Republic; and [f]Division of Plant Breeding and Applied Ge Freising, Germany

Edited by James A. Birchler, University of Missouri, Columbia, MO, and approved July 6, 2012 (received fo

Supernumerary B chromosomes are optional additions to the basic set of A chromosomes, and occur in all eukaryotic groups. They differ from the basic complement in morphology, pairing behavior, and inheritance and are not required for normal growth and

probably a rare event; variants within related gle origin.

One of the best-stu

nature genetics

ARTICL

https://doi.org/10.1038/s41588-019-04

OP

A reference genome for pea provides insight int

legume genome evolution

Jonathan Kreplak[1,20], Mohammed-Amin Madoui[2,20], Petr Cápal[3], Petr Novák[4], Karine Labadie[5], Grégoire Aubert[1], Philipp E. Bayer[6], Krishna K. Gali[7], Robert A. Syme[8], Dorrie Main[9], Anthony Klein[1], Aurélie Bérard[10], Iva Vrbová[4], Cyril Fournier[1], Leo d'Agata[5], Caroline Belser[5], Wahiba Berrabah[5], Helena Toegelová[3], Zbyněk Milec[3], Jan Vrána[3], HueyTyng Lee[6,19], Ayité Kougbeadjo[1], Morgane Térézol[1], Cécile Huneau[11], Chala J. Turo[12], Nacer Mohellibi[13], Pavel Neumann[4], Matthieu Falque[14], Karine Gallardo[1], Rebecca McGee[15], Bunyamin Tar'an[7], Abdelhafid Bendahmane[16], Jean-Marc Aury[5],

ARTICLES

nature genetics

OPEN

A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution

Massimo Iorizzo[1,12], Shelby Ellison[1], Douglas Senalik[1,2], Peng Zeng[3], Pimchanok Satapoomin[1], Jiaying Huang[3], Megan Bowman[4], Marina Iovene[5], Walter Sanseverino[6], Pablo Cavagnaro[7,8], Mehtap Yildiz[9], Alicja Macko-Podgórni[10], Emilia Moranska[10], Ewa Grzebelus[10], Dariusz Grzebelus[10], Hamid Ashrafi[11,12], Zhijun Zheng[3], Shifeng Cheng[3], David Spooner[1,2], Allen Van Deynze[11] & Philipp Simon[1,2]

We report a high-quality chromosome-scale assembly and analysis of the carrot (*Daucus carota*) genome, the first sequenced genome to include a comparative evolutionary analysis among members of the euasterid II clade. We characterized two new polyploidization events, both occurring after the divergence of carrot from members of the Asterales order, clarifying the evolutionary scenario before and after radiation of the two main asterid clades. Large- and small-scale lineage-specific duplications have contributed to the expansion of gene families, including those with roles in flowering time, defense response, flavor, and pigment accumulation. We identified a candidate gene, DCAR_032551, that conditions carotenoid accumulation (*Y*) in carrot taproot and is coexpressed with several isoprenoid biosynthetic

Repetitive DNA characterization using RepeatExplorer

Plants

- Over 100 species characterized so far
- Comparative studies
- Whole genome assembly projects

Mammals

Bats, deer

Fish

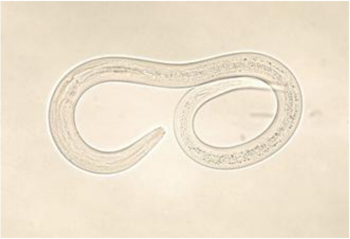
- *Austrolebias charrua, Cynopoecilus melanotaenia*

Insects

- Locust, grasshoppers, kissing bugs

Worms

- Soil helminths



Make you tools public

Galaxy PROJECT

Galaxy Tool Shed

useGalaxy.*

125+ platforms for using Galaxy

The team

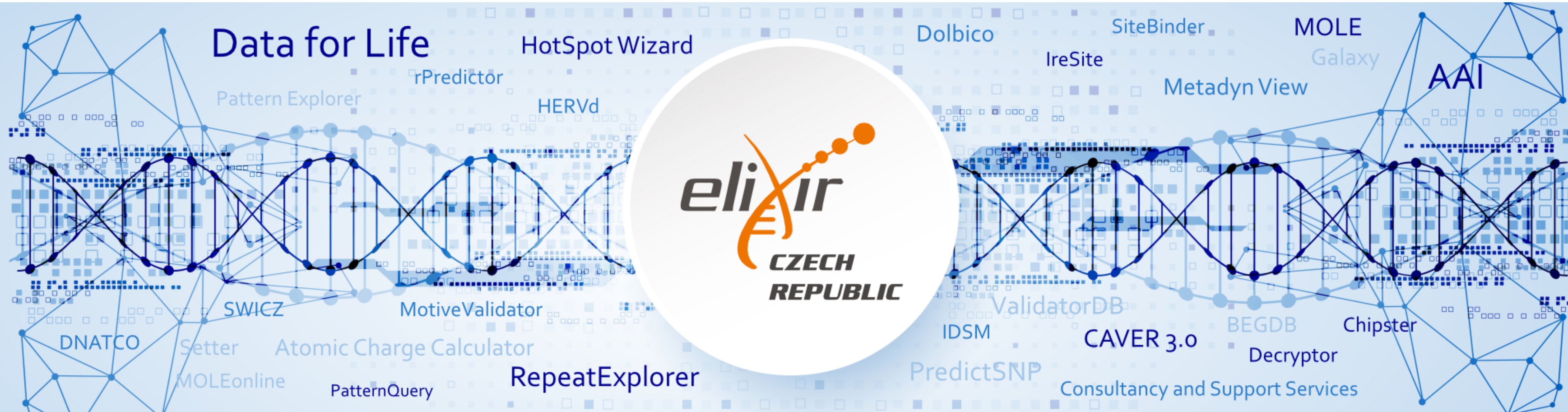
Masaryk Univ. Brno: M. Macháč, I. Křenková, A. Křenek, Z. Salvet

CERIT / CESNET / ELIXIR (hardware)

Laboratory of Molecular Cytogenetics, Biology Centre, CAS

Pavel Neumann **Jiří Macas** **Petr Novák** **Nina Hoštáková** **Tihana Vondrak**

Thank you


Data for Life
HotSpot Wizard
rPredictor
Pattern Explorer
HERVd
Dolbico
SiteBinder
MOLE
IreSite
Galaxy
Metadyn View
AAI
elixir
CZECH REPUBLIC
SWICZ
MotiveValidator
DNATCO
Setter
Atomic Charge Calculator
MOLEonline
PatternQuery
RepeatExplorer
ValidatorDB
IDSM
CAVER 3.0
BEGDB
Chipster
Decryptor
PredictSNP
Consultancy and Support Services

WEBINAR SERIES

Galaxy

Resources for...

Researchers

April 21

Educators & Trainers

April 28

Tool Developers

May 12

Admin & Infrastructure Providers

Gianmauro Cuccuru
Lucille Delisle

May 26

10 am EDT
4 pm CEST